STATE OF CALIFORNIA
DMV
DEPARTMENT OF MOTOR VEHICLES

An Evaluation of California's Commercial Driver License Drive Test

By

Nancy Clarke

May 1995

Research and Development Section
Division of Program and Policy Administration
California Department of Motor Vehicles

(RSS-95-149)

PREFACE

This report is available to interested parties as an internal Research and Development Section document. As such, the findings and opinions contained in the report are those of the author and may not represent the views and policy perspective of the State of California.

ACKNOWLEDGMENTS

The author wishes to acknowledge the individuals who have contributed to this project. The guidance offered by Raymond C. Peck, Chief, Research and Development Section, in the statistical analyses and editing was particularly helpful. Robert Hagge and Rickey Williams, Research Managers, both provided direction for the project. Leonard Marowitz, Research Analyst, wrote a preliminary study proposal. Sharon Seavers, Manager III, acted as liaison between headquarters and the study field offices. Debbie McKenzie, Staff Services Analyst, and Douglas Luong, Office Technician, compiled the final report and provided consistency in the production of tables and drafts.

EXECUTIVE SUMMARY

The federal government requires states which permit third-party testing of commercial drivers to determine whether these tests are equivalent to those given by the state driver licensing authority. To meet this requirement, the California Department of Motor Vehicles (DMV) plans to sample commercial driver license (CDL) applicants tested by their employer, retest them at DMV, and compare the fail rates for the employer and DMV tests to determine if they are equivalent in difficulty and reliability. In order to make this determination, it is necessary to estimate the reliability and other psychometric properties of the California DMV CDL test. Without this information, it is not possible to determine whether differences between the DMV and employer test exceed what would be expected from repeat administration of the CDL test by DMV.

In the present study, a sample of CDL applicants was drawn from 9 of the 10 DMV regions which existed in the state at the time of the sampling. The 131 drivers in the sample were required to complete a test and a retest by DMV.

Total test scores were used to calculate fail rates and mean scores. Results indicate that 36% of drivers in the sample failed the first test, the retest, or both. The percentage of

subjects failing the first test was not computed. However, it probably would have been lower than 36% because some of the subjects passed the first test but failed the second one.

The mean total point score for the total sample of completed tests and retests was 84.3 out of a possible score of 100. For a subsample of completed first tests only, the mean score was slightly lower (83.0). Using analysis of variance (ANOVA) for both the total sample and the subsample, it was determined that there were statistically significant differences among offices on total score ($p<.001$).

The reliability of the test was very low. Interrater reliability was .28 ($p<.05$) and interroute reliability was .51 ($p<.01$). Both of these results are considered inadequate for a drive test. Because net reliability is a joint function of the interrater and interroute reliabilities, the test's net reliability could be lower than the lowest of these two coefficients and would have to be lower than the higher of the two. Chi-square tests were conducted to test for office differences in reliability, and no significant difference on either reliability measure was found.

A Chi-square analysis of fail-rate differences between offices yielded significant results both when DQs were excluded and when they were included. When Chi-square tests were performed on the fail-rate differences between examiners within office, the results were nonsignificant for individual offices or for all offices pooled.

A profile analysis was performed to test whether the field offices had parallel profiles on the 16 scored test items. In contrast to the question of overall differences between offices on total test scores, profile analysis evaluated office differences in the pattern of scores on the individual items. The results confirmed that there were highly significant differences ($p<.001$), indicating possible office differences in scoring procedures and criteria.

The original intent to use the current CDL drive test as a baseline for evaluating third party testing cannot be supported based on the results of this study. The low reliabilities, in particular, underscore the need to consider revising the present CDL drive test into a format similar to the ESSEX model recommended by Mackie et al. (1989). The CDL drive test would then be more consistent with the revised California Class C drive test evaluated by Hagge (1995).

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

INTRODUCTION

Background and Study Objective

For over 30 years, California has participated in third-party testing for commercial driver licenses. Under third-party testing, employers road test their drivers. Those who pass are given a "Certificate of Driving Skill," also known as a "DL 170." The drivers then submit these DL 170s to a Department of Motor Vehicles (DMV) field office for review and issuance of a commercial driver license (CDL).

Title 49 of the Code of Federal Regulations (CFR) addresses CDL standards and Section 383.75 deals specifically with third-party testing. This section requires each state, for purposes of maintaining compliance with testing requirements, either to send government examiners to take the third-party test or to sample employer-certified drivers and compare their fail rates with drivers tested by the state driver licensing agency.

Given these two alternatives for compliance with federal regulation 49 CFR, DMV has chosen to implement the second one by periodically testing samples of employer-certified drivers. However, it was decided that before implementing this option, baseline measures should be developed relative to the following questions: What is the average fail rate for DMV-tested commercial drivers? How reliable is DMV's CDL drive test?

The purpose of this study was to answer these questions for the current CDL drive test. Determining the reliability of California's CDL drive test is important because it will enable DMV to assess more accurately the comparability of employer-certified commercial drive tests with DMV's CDL test. The reliability of the test imposes an upper limit on the extent to which another test can correlate with it. Unless this index is known, it is not possible to judge the performance of a parallel test designed to measure the same attribute. The reliability of a test also imposes a limit on a test's validity. Although a reliable test does not guarantee validity, reliability is a necessary condition for validity. An unreliable test cannot be valid.

This study did not address the issue of test validity per se. Validity deals with whether or not a test measures what it is supposed to measure. Concepts of test validity were addressed by Essex (Mackie et al., 1989) in its development of a prototype CDL drive test.

The present study was conducted to assess the current California CDL drive test. Subsequent to the completion of the study, DMV decided to convert California's CDL drive test to a format which more closely follows the ESSEX model (described below). If this conversion occurs, data from the present study will be used to assess whether the ESSEX-based drive test is more or less reliable than the current CDL drive test.

Literature Review

The following presents a very brief overview of the literature on the reliability of various commercial road tests as reviewed in Peck (in preparation).

The United States Commercial Motor Vehicle Safety Act was enacted in 1986. One of the mandates of this act was the requirement that states have a commercial licensing program which conformed to specific standards of validity and reliability.

Before 1986, the following important studies had been conducted to identify the necessary content domains for commercial road tests.

COMDAT. Engel and Townsend (1984) conducted a study in which they developed and evaluated the Commercial Driver Tractor-Trailer Driving Ability Test (COMDAT). Like the current California CDL road test, the COMDAT examiners scored throughout the test rather than at designated locations along the route chosen a priori. This means that the COMDAT required the examiners to observe and score all behaviors and task elements associated with a maneuver throughout the test.

The internal reliability of the road test was .96. Test-retest administrations of the test over different routes and by different examiners yielded a net reliability of .43 and an interrater reliability of .65.

TORT, TOST, and PTI. McKnight, Kelsey, and Edwards (1984) developed a road test (TORT), an off-road test (TOST), and a pretrip vehicle inspection test (PTI) for testing commercial driver license applicants. The study subjects were recent trucking school graduates.

The interrater reliability for TOST was .93 (N = 47). The interrater reliability for the subsections of the test ranged from .72-1.00. Interroute correlation was considerably lower at .40.

The TORT evaluation involved 373 applicants for a license similar to California's Class A license and 176 applicants for a license similar to California's Class B license. (In California, a Class A license allows the licensee to drive tractor-trailer vehicle combinations. A Class B license permits the driving of buses, fire trucks, and commercial vans.) The investigators randomly assigned subjects to either the experimental group, which took the TORT, or to the control group, which took the regular DMV test. Both tests were given over two routes. The reliability of test scores on the two routes for Class-A-type applicants was .37 for the TORT group and .44 for the control group. For Class-B-type applicants, both the experimental and the control groups had interroute reliabilities near zero.

McKnight et al. reasoned that the low reliabilities were due to the lack of skill variance among commercial driver license applicants, which they hypothesize would be less among commercial applicants than among drivers in general. However, this factor would not explain the failure of TORT to produce more reliable measurements than did the standard California commercial road test.

ESSEX. The American Association of Motor Vehicle Administrators contracted with the ESSEX Corporation for the development of commercial driver licensing test prototypes (Mackie et al., 1989).

The study developed and evaluated a pretrip vehicle inspection, an off-road skill test, and a road test. The road test concentrated on traffic search, direction control, and speed control as the most important driving maneuvers. Its interrater reliabilities ranged from .90-.96, and its test-retest reliabilities varied from .87-.90. The pretrip inspection had interrater and split-half reliabilities of .90, and the skill test yielded interrater reliabilities above .90 for both classes of license applicants. Test-retest reliabilities ranged from .76-.87 for the Class-A-type applicants and from .48-.51 for the Class-B-type applicants.

Although California DMV's commercial drive test has a pretrip inspection, a skill test, and a road test, its road test does not include several important features embodied in the ESSEX model. The above reliabilities, therefore, would not necessarily be expected to be representative of California's CDL road test.

METHODS

Research Design

Field offices for the present study were selected based on representation of geographical location and CDL test volume. One office was chosen from each of the 10 regions, except Region VII which had only one CDL office. Two comparable CDL drive test routes were developed at each of these field offices. Drive test assignments were random, meaning that each subject had an equal chance of being assigned to a specific drive test treatment. The order of examiners and drive test routes comprised the drive test treatments. A drive test treatment consisted of either the same examiner on route 1 and route 2, or two different examiners on the same route. Research and Development created nine drive test treatment matrices. (An example matrix is presented in Appendix A.) These matrices were randomly assigned to the study offices.

Each field office was to test 16 drivers, using two routes and two LREs as follows:

- 4 drivers on route 1 both times with a different LRE each time.
- 4 drivers on route 2 both times with a different LRE each time.
- 4 drivers on routes 1 and 2 with LRE 1 both times.
- 4 drivers on routes 1 and 2 with LRE 2 both times.

Two offices had fewer than 16 drivers due to data collection problems. The test and retest scores obtained in all nine offices were used to evaluate interroute and interrater reliability. Net reliability, which is evaluated by varying both route and examiner, could not be assessed because this would have required having two examiners in the vehicle at the same time. This dual-rater assessment was not conducted due to concern about possible insurance restrictions and the fact that some commercial vehicles are not set up to accommodate having two examiners in the vehicle.

Score sheets were not collected for subjects who failed the pretrip inspection or who were automatically disqualified on the first or second road test. Pretrip inspection failures occurred when drivers were unable to correctly identify and operate crucial vehicle equipment. DQs occurred when drivers committed a maneuver that was so dangerous the examiner ended the drive test immediately. It was decided not to collect detailed test results for these subjects because it was necessary to have two completed score sheets for each subject in order to compute reliability statistics for total point score. Unfortunately, the minimal amount of data that was collected on DQ subjects made it impossible to determine which of the two tests were DQed. It was possible, however, to determine how many applicants were DQed on either one or the other of the two tests.

The sample of subjects in each office was drawn from first-attempt original applicants for a CDL license. This condition meant that no second- or third-attempt applicants were allowed in the sample. This restriction avoided confounding learning effect of familiarity with the drive test route. Applicants who already had commercial driver licenses and only wanted to add endorsements to drive vehicles with special characteristics (e.g., air brakes and passenger transportation), were also excluded in order to avoid drawing a sample heavily biased toward experienced commercial drivers. Trucking school graduates were included in the same proportions as they are normally tested at that field office, to ensure that novice commercial drivers were not disproportionately represented in the study.

Each subject was asked to complete a driver information questionnaire (shown in Appendix B) prior to being tested. Questions covered demographic variables, driving experience, and type of vehicle.

<u>Data Analysis</u>

Descriptive statistics were computed for the following variables: learning method, driver license (DL) class, vehicle type, number of prior accidents and traffic citations, age, gender, and years of driving experience.

Test fail rate and mean total test score by field office and examiner were computed for the total sample of completed tests and retests, and also for a subsample of completed first tests only. The percentage of all subjects in the study who failed one or both of the tests due to a failing total point score or who DQed on either test was also computed. (Subjects could not DQ on both tests, because a DQ on the first test caused a subject to be dropped from the study. It was decided that testing a subject who DQed would have exposed the examiner and the motoring public to unacceptable traffic safety risks.)

Analysis of variance (ANOVA) was performed to assess whether field offices differed significantly on mean total test score for completed tests and retests.

Correlations of test and retest scores (between examiners and between routes) were computed to assess the test's interrater and interroute reliabilities. Office differences on the reliability measures were assessed by performing Chi-square tests on the correlation coefficients converted to Z-Scores using Fisher's r-to-Z transformations. Chi-square tests were also performed to assess whether fail rates differed significantly between study examiners within offices. The latter Chi-square tests were applied to fail rates with DQs included and also to fail rates with DQs excluded.

Finally, profile analysis was used to test whether the profiles of item scores across field offices were parallel. An analysis of item-score profiles is concerned with variations in the magnitude of item-score differences by office across the various test items. This analysis contrasts with those mentioned above, which concentrated on differences in total point score or failure rate on the test as a whole.

The mainframe version of SPSS-X (Statistical Package for the Social Sciences, release 4.1) was the computer software package used for data analysis.

RESULTS

Sample Description

A total of 131 driver information questionnaires were collected. The responses indicated that subjects learned how to drive commercial vehicles from a friend (37.4%) or taught themselves (43.5%), with many saying they employed both methods of learning. The sample was overwhelmingly male (92.4%). The modal age group was 31-54, which included 55.0% of all subjects. Most of the subjects (63.0%) were Class B license applicants. The most common vehicle types operated by subjects were tractor/trailer (38.2%), van (21.0%), and truck (16.0%). The amount of commercial driving experience was quite limited, the average being only 3.6 years. The 13 subjects who were trucking school graduates (none of whom had any commercial driving experience prior to enrollment) had an average of 15.5 years of non-commercial driving experience. Finally, 21.4% of the drivers reported having had one or more traffic convictions, and only 9.9% reported having had one or more accidents. (No information was provided on time period or type of vehicle involved in the traffic incident.)

Test Difficulty

Total tests and retests. Table 1 presents test frequency, fail rate, and mean total test score by field office for the total sample of tests and retests, excluding DQs and pretrip inspection failures.

Table 1

Test Frequency, Fail Rate, and Mean Total Score
by Field Office for Completed Tests and Retests

Field office	Number of completed tests and retests	Fail rate	Mean score (100 maximum)
Total	262	.09	84.3
A	26	.15	76.8
B	32	.12	84.6
C	32	.03	84.6
D	12	.25	81.3
E	32	.00	92.6
F	32	.06	86.0
G	32	.25	76.9
H	32	.00	88.2
I	32	.03	84.8

Note. Results reflect a pooling of tests and retests. All tests for subjects who DQed on either the test or retest were excluded from the computations. The differences between offices were statistically significant on fail rate ($\chi^2 = 24.84$, $df = 8$, $p<.005$) and mean score ($F = 12.6$, $p< .001$).

The average office fail rate was .09 (9%). This fail rate appears to be extraordinarily low. However, one must take into account that DQs and pretrip inspection failures, which, for reasons stated previously, are not reflected in these figures, usually account for most CDL drive test failures. Because the fail rate computation included scores only for subjects who completed both the test and retest, the exclusion of DQs and pretrip test failures (on either of the two tests) obviously deflated the fail rate substantially.

The mean total score on the completed tests and retests for the sample of nine offices was 84.3. The offices differed significantly on this measure (F = 12.6, p<.001, eta^2 = .30), their mean scores ranging from 76.8 to 92.6. (Eta^2 represents the proportion of test score variance that can be attributed to differences among the nine offices, in this case .30 or 30%.)

A better estimate of the difficulty of the test in an actual operational setting is the fail rate of study subjects who failed the test or retest because of points being deducted or for making disqualifying maneuvers. When DQs were included, the total number of subjects was 174. Of these 174 drivers, 63 or 36% failed on points or were disqualified. The reader is cautioned that this .36 fail rate is not comparable to the .09 fail rate, because the former is per 100 drivers and the latter is per 100 total tests (including retests).

Table 2 shows a breakdown of this measure by field office, and also indicates the number of DQs and point-score failures reflected in the results. As expected, there was a very wide range of fail rates across field offices when subject was the unit of analysis and subjects who DQed were included. Fail rates ranged from .11 to .72.

Table 2

Subject, DQ, and Point-Failure Frequencies, and Fail Rate, by Field Office for Subjects Passing the Pretrip Inspection

Field office	Number of subjects	Number of subjects who DQed on test or retest	Number of subjects who failed the test or retest on point score	Fail rate
Total	174	43	20	.36
A	16	3	3	.37
B	16	0	3	.19
C	20	4	1	.25
D	8	2	3	.62
E	18	2	0	.11
F	17	1	2	.18
G	32	16	7	.72
H	28	12	0	.43
I	19	3	1	.21

Note. Results are based on subject as the unit of analysis. Fail rate reflects the proportion of subjects who failed either the first test or the retest, or both, due to a failing total point score plus subjects who DQed on either the first or second test. The differences between offices on fail rate were statistically significant (χ^2 = 33.19, df = 8, p<.001).

First tests only. Table 3 presents test frequency, fail rate, and mean total test score by field office for completed first tests only.

Office test fail rate for the subsample was .14 and ranged from .00 to .47. Again, the fail rates are deflated due to the exclusion of automatic DQs and pretrip failures from the computation. It was not possible to compute a first-test fail rate for all study subjects (i.e., with DQs included), because information was unavailable on which of the two tests resulted in a DQ rating.

The mean total test score for all completed first-test scores was 83.0. Mean scores ranged from 74.4 to 92.4. The differences between field offices on this measure were statistically significant ($F = 7.7$, $p<.001$, $eta^2 = .33$). This result could reflect differences in office scoring procedures, differences in applicant skill, or both.

Table 3

Test Frequency, Fail Rate, and Mean Total Score
by Field Office for Completed First Tests Only

Field office	Number of completed first tests	Fail rate	Mean score (100 maximum)
Total	133	.14	83.0
A	14	.14	75.4
B	16	.19	82.8
C	16	.00	84.6
D	8	.37	76.4
E	16	.00	92.4
F	16	.12	82.8
G	15	.47	74.4
H	16	.00	88.4
I	16	.06	84.6

Note. Tests taken by subjects who were DQed on either the first test or the retest were excluded from the computations. The exclusion was necessary because information needed for determining which of the two tests were DQed was not collected. The differences between offices were not statistically significant on fail rate ($\chi^2 = 25.88$, $df = 8$, $p< .005$) or mean score ($F = 7.7$, $p< .001$).

Consistency in Examiner Scoring

Test reliability. Tables 4 and 5 present the test-retest interrater and interroute reliabilities. Interrater reliability for the sample of nine offices pooled was .28, which is statistically significant at $p<.05$. This figure was derived by pooling office results and correlating the scores given by one examiner with the score given by the second examiner on the same route. This extremely low reliability value means that 72% ([1.00 - .28] x 100%) of the variance in total test scores was due to differences in examiner scoring or measurement/sampling error. Although four of the offices appear to have had inverse scoring relationships between the examiners (i.e., negative coefficients), the differences between the obtained reliabilities of each office did not approach significance ($\chi^2 = 6.44$, $df = 7$, $p = .50$). The fact that the very large differences in observed reliabilities across office were not significant underscores the instability of the estimates within each office, which are based on very small sample sizes.

Table 4

Number of Test-Retest Pairs, Interrater Reliability, and Mean Examiner Scores by Field Office

Field office	Number of test-retest pairs	Interrater reliability	LRE 1 mean score	LRE 2 mean score
Total (pooled)	65	.28	82.3	83.8
A	6	.17	68.0	76.3
B	8	.67	83.3	81.4
C	8	.22	85.0	81.0
D	3	-.94	79.7	65.7
E	8	.36	92.8	91.6
F	8	-.34	80.8	92.6
G	8	.39	72.9	76.4
H	8	-.31	87.8	89.8
I	8	.30	85.4	85.9

Note. Reliabilities represent the pooled correlations of LRE 1 and LRE 2 scores for routes 1 and 2. Computations were based on completed tests and retests only (i.e., scores for subjects who DQed either test were excluded). After converting correlations to *Z*-scores using Fisher's transformation (Snedecor & Cochran, 1967), it was determined that the differences between offices were not statistically significant ($\chi^2 = 6.14$, $df = 7$, $p = .50$). Because office D had only three test-retest pairs, it was dropped from the Chi-square analysis.

Table 5

Number of Test-Retest Pairs, Interroute Reliability, and Mean Route Scores by Field Office

Field office	Number of test-retest pairs	Interroute reliability	Route 1 mean score	Route 2 mean score
Total (pooled)	66	.51	84.8	86.4
A	7	.48	77.7	83.7
B	8	.38	85.1	88.3
C	8	.64	86.0	86.4
D	3	.09	88.0	92.0
E	8	.32	94.0	92.0
F	8	.30	82.8	87.9
G	8	.48	79.3	79.0
H	8	.68	87.4	87.8
I	8	.44	83.8	84.0

Note. Reliabilities represent the pooled correlations of route 1 and route 2 scores for LRE 1 and LRE 2. Computations were based on completed tests and retests only (scores for subjects who DQed either test having been excluded). After converting correlations to *Z*-scores using Fisher's transformation (Snedecor & Cochran, 1967), it was determined that the differences between offices were not statistically significant ($\chi^2 = 1.41$, $df = 7$, $p = .98$). Because office D had only three test-retest pairs, it was dropped from the Chi-square analysis.

Interroute correlation was .51, which is statistically significant at $p<.01$. This result represented the correlation of route 1 scores with route 2 scores for the two examiners. Although this result was more acceptable than the interrater reliability, it still indicated that 49% of the variance in total scores was due to interroute differences or measurement/sampling error. Again, the variation in reliabilities across offices was not significant ($\chi^2 = 1.41$, $df = 7$, $p = .98$).

One would expect scores on the two routes to be very similar, since field offices are required to have alternate parallel routes, which are created using specific guidelines. Even if the routes are perfectly parallel, however, it is possible that their length and content is not sufficient to cope with the degree of within-subject variability over the route conditions and sequences. What is more surprising is that interrater correlation is so low. Both of these reliability estimates are much lower than those for the California Class C drive tests (Shumaker, 1994; Hagge, 1994).

It should be noted that net reliability, which was not evaluated, could even be lower than .28 and would have to be lower than the interroute reliability estimate of .51. This follows from the fact that net reliability is a joint function of the two error components (rater and route variance).

Fail rate differences. Chi-square tests performed on fail rates across offices with DQs excluded, as reported in Table 1, revealed significant differences among offices ($\chi^2 = 24.84$, $df = 8$, $p<.005$). Chi-square tests were also performed on the fail rates with DQs included (Table 2) and the results were similar ($\chi^2 = 33.19$, $df = 8$, $p<.001$). The same general pattern emerged when the analysis is confined to first tests only, as shown in Table 3 ($\chi^2 = 25.88$, $df = 8$, $p = .005$). It is therefore clear from these three analyses that test failure rate varied as a function of field office. It is not possible to determine whether these differences reflect differences in applicant skill level or office scoring standards. However, the effects of differences between examiners can be isolated by comparing LREs within each office (Table 6). This analysis indicated that none of the differences between examiners within office was significant, as evidenced by the Chi-square results for individual offices ($df = 1$). The additive property of the Chi-square statistic allows the combining of the individual Chi-squares into a more powerful single test of differences (Hedges & Olkin, 1985). The pooled Chi-square result was highly nonsignificant ($\chi^2 = 2.16$, $df = 7$, $p = .95$). Thus, when differences between offices are controlled, there is no evidence of differences in LRE failure rates.

Table 6

LRE Fail Rate Difference by Field Office
Using Yates' Continuity Correction

Field office	Number of tests and retests	Fail rate difference	χ^2	p
A	26	.15	0.30	.59
B	32	.06	0.00	.99
C	32	.06	0.00	.99
D	12	.56	1.33	.25
F	32	.12	0.53	.47
G	32	.00	0.00	.99
I	32	.06	0.00	.99

Note. Chi-square could not be computed for offices E and H because these two offices had no point-score failures. DQs were excluded from the computations. Total pooled Chi-square is nonsignificant ($\chi^2 = 2.16$, $df = 7$, $p = .95$).

Profile analysis. For this analysis, the drive test was divided into 16 components. The purpose of the profile analysis was to determine whether the nine field offices had parallel patterns of scoring on the 16 components. The analysis yielded significant results ($F = 3.05$, $p<.001$, $eta^2 = .94$), indicating that the component profiles were not parallel. This means that the relative magnitude of item means within the total test varied across offices.

The analysis also tested whether group effects were equal. Did the same average response occur from field office to field office? The results were significant ($F = 7.55$, $p<.001$), indicating that there were significant differences between the offices on the component and the total test scores.

Finally, profile analysis tested for equality or "flatness" of the component means. (Were the test segments equal to each other across field offices?) This effect is of little interest since one would not expect the items to be of equal difficulty, which they were not.

DISCUSSION AND CONCLUSIONS

The mean total point score for the total sample of completed tests and retests was 84.3 out of a possible score of 100. There were significant differences among field offices on this measure ($p<.001$). These results indicate that there were real differences between offices in applicant performances, examiner scoring, or both.

The point-score fail rate (.09) for the pool of tests and retests was extremely low, as would be expected considering that DQs and pretrip inspections were excluded from its computation. When DQs and all tests were included in the analysis, the fail rate increased substantially, to .36.

The offices also differed significantly on fail rate ($p<.01$). This was true irrespective of whether or not DQs and retests were included in the analyses. A Chi-square analysis of fail-rate differences between examiners, however, showed nonsignificant results. Thus, there is no evidence of differences in failure rates between LRE's when the effects of field office are controlled. The significant differences between the test scores for the offices could reflect differences in applicant skill level and/or differences in office scoring standards.

Interrater reliability was very low (.28), but interroute reliability was somewhat higher (.51). Both types of reliability, especially interrater, can probably be improved through refresher training for drive test examiners. The inadequacy of these test reliability results becomes more apparent upon recognizing that the total net reliability of the test would have to be lower than .51 and could possibly be lower than .28. The low interrater and interroute reliabilities underscore the need to consider revising the current CDL test along the lines recommended in the original ESSEX study (Mackie et al.). This would also yield a test that is more consistent with the recent revisions in the California Class C drive test, as described in Hagge (1994). In any event, the original objective of using the present CDL tests as a baseline for evaluating third party testing cannot be supported based on the results presented here.

REFERENCES

Engel, R., & Townsend, M. (1984). *Commercial driver tractor-trailer driving ability test manual*. Ottawa, Ontario: Transport Canada.

Hagge, R. (1994). *Class C driving performance evaluation: Stage 3 study*. Sacramento, CA: Department of Motor Vehicles.

Hedges, L. & Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin*, 96 (3), 573-580.

Mackie, R. R. Wylie, C. D. Shultz, T. Engel, R. Townsend, M. Lammilien, S. E. and Johnson, S. (1989). *Development of a recommended testing program for commercial motor vehicle operators (The CDL system), final report.* The Essex Corp.; Engel and Townsend; and Personnel Decisions Research Institute.

McKnight, A. J., Kelsey, S. L., & Edwards M. L. (1984). *Development of knowledge and performance tests for heavy vehicle operators: Volume I development and field test.* Sacramento, CA: Department of Motor Vehicles.

Peck, R. (In Preparation). *Driver licensings and highway safety*. Sacramento, CA: Department of Motor Vehicles.

Shumaker, N. (1994). *The California driver performance evaluation project: An evaluation of the current driver licensing road test.* Sacramento, CA: Department of Motor Vehicles.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods (6th ed.)*. Ames, IA: Iowa State University.

SPSS Inc. (1988). *SPSS-X user's guide (3rd ed.)*. Chicago, IL: Author.

Tabachnik, B., & Fidell, L. (1989). *Using multivariate statistics (2nd ed.)* New York: Harper & Row.

APPENDIX A

Example of Drive Test Schedule

Routes (T = test; R = retest)

Driver	Route 1		Route 2	
	LRE 1	LRE 2	LRE 1	LRE 2
1			R	T
2	T		R	
3	R	T		
4	R		T	
5	R		T	
6			T	R
7	T	R		
8		T		R
9		R		T
10	T		R	
11			T	R
12	T	R		
13	R	T		
14		T		R
15		R		T
16			R	T

APPENDIX B

Commercial Driver's License Drive Test
Evaluation Driver Information Form

1. Name last_____________________first______________________middle______________

2. Birthdate________________________ 3. Age_________

4. Address street___________________city___________________ZIP CODE_____________

5. Type of vehicle________bus________tractor/trailer_______bobtail truck (check one)

6. Current driver license #____________________

7. Out-of-state application?_____yes______no

8. Please answer these questions about your driving experience. (Information is confidential, for research only, and will not become part of your driving record.)

   What is the name of the trucking firm which has or is going to hire you?

   ____

   How many years of commercial driving experience do you have?_______years

   Do you have any prior traffic convictions or accidents? _____yes _____no

   If answer is "no," go to question 9.

   If answer is "yes," please answer these questions.

   # traffic convictions________ # accidents________

9. Are you a recent trucking school graduate?_______yes______no

   If answer is "no" go to question 10.

   If answer is "yes" please answer these questions. (Information is confidential, for research only, and will not become part of your driving record.)

   Name of trucking school____

   How many years of experience do you have driving with a non-commercial license?____years

   Do you have any prior traffic convictions or accidents?_____yes_____no

   If answer is "no" go to question 10.

   If answer is "yes," please answer these questions.

   # traffic convictions_________# accidents_________

10. How did you get your commercial driving experience?

    a. taught self _________
    b. learned from a friend _________
    c. previously licensed as a commercial driver _________